UIS Informatics Commission

Publications Exchange Working Group

Chairman: Trevor FAULKNER, United Kingdom

Appendix 4: Scanning documents for digital archiving

Issue 2: 9 November 2016

This discussion covers making electronic copies of paper publications. Digital archiving of photographs, movie films, video tapes, and sound recordings is beyond its scope. Some of it will apply to survey notebooks and other personal documents. The reasons for scanning publications that exist only on paper can range from just having a backup version that can be stored in several places away from the original as protection against loss, to preparing a version that can be viewed on-screen or printed that is as good as or perhaps better than the original. In the first case, just acquiring files of raw scans that preserve all the information on the original pages and storing them, properly cataloged and dispersed, can suffice. Additional effort is needed to turn the raw scans into a “reprint” suitable for printing a facsimile of the original or convenient reading on a computer screen. Additional benefits can be gained if the resulting file is put through the optical-character-recognition process (OCR), so that the text can be searched and, if appropriate, indexed by services such as Google.

The goal when making the raw scans is to acquire grayscale, or colour if necessary, scans with a resolution of 300 or 400 DPI. If the original page contains very small print or a map or other illustration with very fine detail, a resolution of 600 DPI may be needed; resolutions higher than that are unlikely to be useful because of the texture of the paper or the inherent resolution of the original printing technology. There is no need for more accuracy in levels than 8-bit grayscale or 24-bit color. If the original is extremely clear and contains only black text or line drawings, a black-and-white scan (1 bit per pixel) might suffice, but that risks losing information that might be useful during further processing of the scan. (While black-and-white, as opposed to colour, sometimes encompasses grayscale illustrations, here it is distinct from them.)

It is easiest to obtain raw scans that are not distorted if the publication can be disbound for scanning, either by just removing staples or taking apart the book. Staples can of course by easily replaced, and rebinding, though probably not of the original quality, can be done if necessary. Of course whether this is done depends on the balance between the intended use of the scan and the value of the original, which may be old or rare.

The most common devices for scanning are flatbed scanners that handle pages up to A4 or US letter in size. This size will accommodate most pages of most publications. If paper is printed on both sides, some enhancement in quality of the scans may be gained by backing up the sheet being scanned with a sheet of black paper. This will keep the material on the back from showing through on the scan of the front if the paper is thin. Some multi-function printer/copiers include scanners that sheets are fed through, sometimes automatically from a stack. Needless to say, the latter can be used only for loose pages in good condition.

Pages that must be scanned from bound volumes can still be scanned on a flatbed scanner, but distortion near the binding edge is likely unless the inside margin was generous; pressing down to reduce this can damage a valuable volume. Some libraries have overhead book scanners, which scan from above a book laying open on the table. If available, this is a good way to scan a bound volume, although distortion where the page curves into the binding will still be visible. It might be possible to construct something similar by using a digital camera, or perhaps even a smart-phone, to acquire the image. Hand-held scanners exist that can be moved across, rather than down, the page. These might make it possible to reduce the distortion at the binding.

The Karst Information Portal (www.karstportal.org) offers the service of scanning originals for its archive, and provides a copy of the result. The KIP runs on a Cloud-hosted Drupal platform. It seemed better for Portal managers to control the data rather than to use the dSpace contributor support model. The KIP collection has steadily increased the number of contained digital objects. In June of 2006, it hosted just over 3,000 metadata records without digital content. Today it hosts nearly 7,000 metadata records, 43 percent of which link to digital content included in the KIP collection. For questions or comments concerning the KIP project, please contact Todd Chavez.

Foldouts or large maps that were originally distributed with the publication in loose, folded form will have to be dealt with separately. Foldouts can probably just be scanned in two or three pieces using the same equipment as the rest of the book, or a hand-held scanner could scan them all at once. Large sheets can also be scanned in pieces, but it is much easier to have them scanned at a library, copy shop, or similar service that has a large scanner. Whether the pieces need to be reassembled by computer depends on the intended use. For just archiving, the considerable effort involved may not be needed.

It is critically important to prevent inappropriate compression of the raw scans. Some compression schemes, in particular, the popular JPEG, are designed only for photographs and obtain their impressive compression ratios by eliminating detail that the eye will not perceive. When applied to text or line drawings, however, the algorithm degrades the image in ways that can easily be seen, and JPEG compression is not reversible; the original cannot be recreated from a JPEG (.jpg) file. The best way of storing image files is as TIFF (.tif) files. They can be compressed reversibly without loss of information using ZIP or LZW compression schemes. Essentially all programs that can read or write image files can deal with TIFF files. Most can deal with those types of compression, which still leave the file as type .tif. Make sure the program you used to scan is not set to save JPEG files.

Whatever the choice of digital medium, which is beyond the scope of this section, it will be most convenient to package raw scans into a single file for the publication, or maybe even a string of issues that make up a volume, rather than store perhaps hundreds of files. A good way to do this is to convert them all to PDF files and concatenate all the PDFs into a single PDF file. Some programs can do the conversion and concatenation in one step. (Some scanning programs will create the PDFs themselves, but might not give control over compression, and they will not be easy to manipulate further.) PDF is just as standard a format as TIFF and not likely to become obsolete any sooner, and if the PDF is properly made, the scan images can be easily copied back out of it if necessary. Again, care is needed about JPEG compression; some programs that create PDFs use it for grayscale or color images by default. Be sure that the creation of the PDF does not resample the file to lower DPI. Experiment with the programs in use, to make sure the images can be extracted, and that the extracted images do not show JPEG artifacts and that they have the same DPI as the scans. Be sure to include loose maps, either the pieces or as oversize pages, at the end of the PDF. An alternative is to package all the scans in their folder into a single .zip file by compressing the folder. Even if the intended use of the scans is to produce a good "reprint," the raw scans should be kept, although in that case redundancy of storage may not be important once the final product has been distributed. If the scanned publication is to be made widely available in libraries or on a web site, it is best to make the "reprint" attractive and as close to the original as possible. In this case a number of additional steps will be needed. Just what is done will depend on the software and time available. [Insert here a reference to my detailed write-up on manipulating scans if it is included in the report.]

The raw scans are likely to be crooked or contain blemishes beyond the edges of the printed page. Straighten up the image and crop to include just the printed area. The alignment should be done as accurately as possible. The cropping needn't be tight. Type and line art such as maps that are strictly in black-and-white will occupy less file space and print more sharply if converted to black-and-white from grayscale or color. It may be necessary to adjust the density levels in the raw scans to get an optimum conversion. If the quality of the original is very poor, it may be impossible to find an adjustment that leaves the text legible after conversion. Leave it grayscale.

If a page has grayscale or color illustrations, leave the whole page grayscale or color, as scanned. However, to get the best results for all the parts of a page, it is possible to remove the grayscale or color images into separate files and optimize them for contrast and color separately; the remainder of the page can then be converted to black-and-white. If the illustrations were originally printed poorly, it may be possible at this stage to improve them, not just reproduce the original. When printed, such illustrations will have been half-toned, i.e. converted to black-and-white or colored dots. Half-toned photographs can be improved by "descreening," blurring out the dots, without affecting the perceived sharpness. Image-editing software usually has some filter that will accomplish that. Avoid doing it to maps or charts, as there the resulting blurring may be worse than leaving the dots. Some scanning software offers to de-screen, but do not do that, because it may end up compromising the raw scan in undesired ways.

Any program that is capable of accepting large image files on a sufficient number of pages can be used to assemble the new version of the publication. Place the cropped black-and-white image of the black-and-white material on a page, duplicating the original margins and location, and then add whatever illustrations have been removed to be treated separately. It may be necessary to stretch slightly illustrations that were printed to the very edge of the page, if scanning or straightening left them too small. At this stage, make sure that the original pagination of the book or magazine is preserved, with any blank pages, including back sides of covers. This will result in a file that has black-and-white, for sharpness and economy of space, except where grayscale or color is needed.

After scanning, save, export, or whatever it is called by the program, into a PDF file. Again, be careful that grayscale and color illustrations do not get JPEG-compressed, which is sometimes the default in the conversion process, and that the DPI of the images is not changed. Fortunately, there is no such thing as JPEG-compressed black-and-white, so only non-photo grayscale or color illustrations are a potential problem here. (If all the grayscale or color illustrations are photographs, high-quality JPEG compression will do no real harm, but the result will not be absolutely faithful to the original.) Permitting ZIP compression if given a choice will reduce the file size without compromising quality, and it is especially effective in reducing the size of the black-and-white material that probably makes up most of the publication. If in doubt, view the resulting PDF at high magnification to check for JPEG artifacts, which will manifest themselves as fuzz around lines or letters in grayscale or color graphics. If the original publication was large with many illustrations, the result can be a file of several hundred megabytes. Nevertheless, this is what should be kept or distributed to other libraries. For viewing on the web, a much smaller version might be created by allowing aggressive JPEG compression; this can be done from the good PDF, without going back to the layout. The tradeoff between the quality of some illustrations and data transmission time and storage space might be acceptable in that context.

If talent, ambition, and software permit, a nice touch is fixing the page numbers known to the PDF reader to match the page numbers in the document, so that, for instance, page ix is shown in the header on the screen as ix. Bookmarks can also be added for easy access to chapters or articles. Also, as mentioned earlier, it can be very helpful to produce an OCR file of a PDF, so that readers can search for text on-screen and, if it is publicly accessible on the Web, search engines can index it.

This guideline document has been drafted by Bill Mixon for the International Union of Speleology's Publications Exchange Working Group in April 2016. Comments, suggestions for improvement or clarification, and critical input are invited from interested parties. Please address all correspondence to Bill Mixon or the Chairman.